Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Daryna Dementieva

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Methods for fighting with harmful multilingual textual content

Supervisor: Assistant Professor Alexander Panchenko

Name of the Reviewer: Prof. Georg Groh, TU München, Germany

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: Sept 4th, 2022**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense. *If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.*
Reviewer's Report
Reviewers report should contain the following items: • Brief evaluation of the thesis quality and overall structure of the dissertation. • The relevance of the topic of dissertation work to its actual content • The relevance of the methods used in the dissertation • The scientific significance of the results obtained and their compliance with the international level and current state of the art • The relevance of the obtained results to applications (if applicable) • The quality of publications The summary of issues to be addressed before/during the thesis defense

(see attached extra assessment paper)

Provisional Recommendation

X ☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*

Assessment of the PhD-Thesis

Methods for Fighting with Harmful Multilingual Textual Content

submitted to
Skolkova Institute of Science and Technology, Moscow
by Daryna Dementieva, MSc

Research Questions and Subjects of the Thesis

The thesis discusses new multilingual approaches to tackle types of harmful textual information such as fake news and toxic speech. The thesis is submitted in a form of a coherent academic treatise (in TUM terms a cumulative dissertation) consisting of 10 chapters and is based on 11 papers published in scientific venues.
The first two chapters are an introduction and a background chapter. The following 7 chapters cover the material published as publications and a short description of the contribution of the applicant to the work. The final chapter summarizes the scientific contributions of the thesis.

The first part of the Introduction chapter outlines the motivation of the covered research. The part is ended with the statement of the main research questions that the work is dedicated to: "Q1: Can multilingual evidence from external sources improve fake news detection?"
and "Q2: Which NLP technology (both monolingual and multilingual) can be used to detoxify texts?" (cited from dissertation)
The second part is a short description of the content of each chapter emphasizing the contributions made in each one.
The background chapter (basic introductory related work chapter) provides an overview of the topic natural language processing for social good. The part starts with a definition of AI for social good tasks and problems and the criteria of "social good" technologies. After that, examples are provided of which problems can be tackled specifically in the scope of the NLP field. The second section is an overview of the modern state of multilingual NLP technologies and models that will be used in the next parts of the thesis.

The rest of the thesis is divided into two parts corresponding to two types of harmful information covered in the work. The first part is dedicated to the task of fake news detection. Chapter 3 provides motivation and a formal definition of the task. A comprehensive overview of related work that describes the existing datasets and models for fake news detection motivates and

Technische Universität München
Fakultät für Informatik

Prof. Dr.
Georg Groh
Boltzmannstr. 3
85748 Garching

Tel. +49 89 289 18678
Fax +49 89 289 18657

grohg@in.tum.de
www.social.in.tum.de

Bayerische Landesbank
IBAN-Nr.:
DE10700500000000024866
BIC: BYLADEMM
Steuer-Nr.: 143/241/80037
USt-IdNr.: DE811193231

substantiates the gap in multilingual data and methods for the fake news detection field which is subject of the first part of the thesis.
Chapter 4 logically continues the line of the research presenting the main contribution of Part 1 – a new feature based on cross-lingual evidence scraped from a Web search. The hypothesis of such new feature exploration is clarified. There were two experiments confirming the stated hypothesis – a manual check and an automated experiment with fake news detection systems. Both experiments have thorough descriptions. The baselines and several datasets for the test are provided, and the results are clearly stated justifying the proposed methods and solution ideas.
Chapter 5 discusses several new metrics for multilingual and cross-lingual similarity of news texts, multilingual large language models for text embedding, and several approaches to extract facts from texts. The chapter is closed with a demonstration of how a combination of new multilingual news similarity metrics and the fake news detection system can be used to provide explanations to a user about the decision of fake news classification.

Part 2 is dedicated to the detoxification task. This part repeats the concept of a structure of the previous one.
The first chapter in this part is an introductory chapter for the task. It covers task motivation, the formal definition of a text style transfer task, and a description of related work for text style transfer in general and for detoxification tasks in particular. A conclusion of a chapter is a motivation for the solution concepts.
Chapter 7 is a description of ParaDetox – a new method for parallel dataset collection for the detoxification task. The chapter covers the description of all steps of the pipeline and data quality control. Two newly collected datasets – for the English and Russian languages are presented.
New methods for detoxification are discussed and evaluated using the new datasets in Chapter 8. Firstly, condBERT - a new unsupervised approach for text style transfer was discussed. Also several seq2seq models with unsupervised baselines were tested for monolingual English and Russian detoxification. The usage of parallel data allows to achieve high-quality detoxification, which is the main contribution of this part. Additionally, experiments for multilingual and cross-lingual detoxification tasks are discussed. The chapter ends with a discussion of cases where the proposed models can be used. Part 2 ends with a discussion of the correlation between the human and automatic evaluation of the detoxification task.

The conclusion chapter summarizes the contributions of each chapter. The answers to the research question stated in the Introduction chapter are provided. At the end of the work, some discussions are given on what further directions of research can be within the framework of the topics described.

Originality, Relevance, and Definition of the Research Questions, Choice of Research Methodology

Considering the related work discussed in the thesis, the research question(s) of the thesis are not extraordinarily new. However, the author is able to further develop and combine existing ideas in a clever clever way into new approaches / variants of approaches that are able to avoid a number of weaknesses of previous approaches.
All research questions of the thesis are relevant for NLP and social applications for NLP in particular without a question and are well defined.

The research methodology chosen is appropriate: effectively a design-science inspired engineering methodology (creating a solution artefact and evaluating it in relation to other methods on suitable data-sets).

Scientific Rigor, Consideration and Presentation of State-of-the-Art Related Work

Daryna's way of citing related work and own work is without any objections. She includes, presents, relates, and discusses state-of-the-art related work in appropriate depth, recency, and extension.

Clarity and Soundness of the Argumentation, Structuring

The thesis excels not only in terms of the ideas for the actual "solution approaches" but also in the thorough way of evaluating the proposed approaches. This contributes to rooting and relating the solution elements in the body of known approaches and clearly demarcates the scientific advances and knowledge gain.
The logical structuring of the content is also very good and contributes well to the value of the thesis.

Form and Language

The Latex-based thesis is immaculately formatted and well-structured layout-wise. The English is mostly correct, scientifically appropriate, precise, clear, and logical.
The author also manages to compactly visualize and communicate her ideas and results in thoroughly crafted tables and figures.

Publications, Impact and Scientific Relevance of the Results

The thesis is based on 8 first-author papers and 3 second-author papers. 3 articles are published in A/A* conferences, 1 article is published in Q2 journal indexed in WOS/Scopus, and others are published in Scopus indexed conferences and workshops. This securely fulfills the requirements of the PhD thesis defense policy of Skoltech.
The scientific quality of all papers is very good.
The number and quality of venue of the papers can be considered well above average for a good PhD-work, the high impact (30 citations in the last 3 years), and excellent scientific quality of the papers are convincing and top-level compared to PhD theses at the TUM Faculty of Informatics.

Quantity, Quality, and Originality of the Candidate's Scientific Work

Considering the thoroughness, she puts into her work, the quantity of her results in relation to its quality and originality is absolutely sufficient and suitable for a top-level PhD-thesis.
Respecting related work while cleverly and originally advancing from it is how research should be done.

Although a significant part of the results was obtained together with other researchers, Daryna has provided leading contributions to these results.
The scientific creativity demonstrated by the scientific contributions is significantly above average compared to other PhD work in the faculty of Informatics.

Overall Assessment

Daryna Dementieva's PhD thesis and her research works clearly demonstrate her competence for original scientific research. I recommend the acceptance of the thesis and would grade it in the current PhD thesis grading system at TUM with the predicate

with highest distinction (summa cum laude)

The research work of Daryna produced relevant and original scientific results in a depth and extension fully appropriate for a PhD degree.
The solution ideas and their presentation are among the best of PhD candidates at TUM and thus deserve the predicate "with highest distinction". The number of publications, the reputation of the venues, and the high number of citations clearly show the excellent quality, originality, and relevance of her research.

In her thesis and her papers I do not see any signs of plagiarism.

Garching, Sept 4th, 2022